Ilmenauer Beiträge zur Wirtschaftsinformatik

Herausgegeben von U. Bankhofer, V. Nissen
D. Stelzer und S. Straßburger

Steve Röhrig, Tobias Rockel

Analyse existierender Simulationsstudien zum Umgang mit fehlenden qualitativen Daten

Arbeitsbericht Nr. 2020-04, Dezember 2020

Technische Universität Ilmenau
Fakultät für Wirtschaftswissenschaften
Institut für Wirtschaftsinformatik

Autor: Steve Röhrig, Tobias Rockel

Titel: Analyse existierender Simulationsstudien zum Umgang mit fehlenden qualitativen Daten

Ilmenauer Beiträge zur Wirtschaftsinformatik Nr. 2020-04, Technische Universität Ilmenau, 2020

ISSN 1861-9223

ISBN 978-3-938940-63-1

URN urn:nbn:de:gbv:ilm1-2020200439

Anschrift: Technische Universität Ilmenau, Fakultät für Wirtschaftswissenschaften und Medien, Institut für Wirtschaftsinformatik, PF 100565, D-98684 Ilmenau.
http://www.tu-ilmenau.de/wid/forschung/ilmenauer-beitraege-zur-wirtschaftsinformatik/

Gliederung

Zusammenfassung: Das vorliegende Arbeitspapier betrachtet die Struktur von 30 Simulationsstudien, welche die Güte von verschiedenen MD-Verfahren für fehlende Werte in qualitativen Merkmalen untersuchen. Für die Betrachtung wird zunächst der allgemeine Aufbau der Studien beschrieben. Des Weiteren werden Merkmale der in den Studien verwendeten Datenmatrizen erhoben, aggregiert und ausgewertet. Dabei ist auffällig, dass z. B. ordinalskalierte Merkmale verhältnismäßig wenig untersucht werden. Darüber hinaus werden mit den verwendeten Ausfallraten gemeinsam mit den Ausfallmechanismen und -mustern die Variationen der fehlenden Daten in den Studien betrachtet. Ein weiterer Überblick wird außerdem zu den verwendeten MD-Verfahren und Bewertungskriterien gegeben. Dabei ist auffällig, dass bei den MD-Verfahren und den Bewertungskriterien sehr viele Unterschiede zwischen den Studien existieren, wodurch im Endeffekt kein MD-Verfahren oder Bewertungskriterium in mehr als 60 % der Studien verwendet wird.

Schlüsselworte: fehlende Werte, Imputation, missing data, qualitative Daten, Simulationsstudien

1 Einleitung

Die meisten Verfahren zur Datenanalyse setzen eine vollständige Datenmatrix voraus und erschweren somit die Analyse unvollständiger Datenmatrizen (vgl. Schafer und Graham 2002, S. 147). Das Problem der Analyse unvollständiger Daten ist praxisrelevant, da in der Realität viele Datenmatrizen unvollständig sind (vgl. z. B. Backhaus und Blechschmidt 2009, S. 266; Eekhout et al. 2012, S. 729–731). Wenn eine unvollständige Datenmatrix vorliegt, stehen laut Bankhofer (1995, S. 89–90) fünf verschiedene Strategien zur Analyse der Daten zur Verfügung: Eliminierungsverfahren, Imputationsverfahren, Parameterschätzverfahren, Multivariate Analyseverfahren und Sensitivitätsverfahren. Eine ausführliche Darstellung dieser fünf Ansätze ist bei Bankhofer (1995) zu finden.

Die Auswahl einer geeigneten Strategie stützt sich unter anderem auf den vorliegenden Ausfallmechanismus. Das Konzept der Ausfallmechanismen geht auf Rubin (1976) zurück und beschreibt den Zusammenhang zwischen der Datenmatrix und der Missing Data (MD) Indikatormatrix. Letztere gibt an, ob ein Wert in der Datenmatrix beobachtet ist oder nicht. Die Daten werden als Missing at Random (MAR) bezeichnet, sofern das Fehlen der Werte nicht von den unbeobachteten Werten abhängt. Ein Spezialfall von MAR stellt der Missing Completely at Random (MCAR) Ausfallmechanismus dar, bei dem das Fehlen der Werte nicht von den Werten der Datenmatrix abhängt. Falls die Bedingungen für MAR nicht erfüllt sind, wird von einem Missing Not at Random (MNAR) Ausfallmechanismus gesprochen (vgl. Little und Rubin 2020, S. 13–14).

Neben dem Ausfallmechanismus ist das Skalenniveau ein weiteres wichtiges Kriterium für die Auswahl eines geeigneten Verfahrens. In der Literatur ist hierbei die Betrachtung von Verfahren für quantitative Merkmale vorherrschend (vgl. Ferrari et al. 2011, S. 2410; Cugnata und Salini 2017, S. 316). Da jedoch viele Datenmatrizen existieren, die nicht nur aus quantitativen Merkmalen bestehen, fokussiert sich dieses Arbeitspapier auf Methoden zum Umgang mit fehlenden Werten bei qualitativen Merkmalen. Das Ziel dieses Arbeitspapiers ist dabei die Analyse existierender Simulationsstudien zum Vergleich von Methoden zum Umgang mit fehlenden Werten bei qualitativen Merkmalen. Ähnlich wie bei Lin und Tsai (2020) wird in diesem Arbeitspapier der Aufbau solcher Studien analysiert, um so Rückschlüsse über bereits existierende Vergleichsansätze und eventuell vorhandene Forschungslücken ziehen zu können.

Für das vorliegende Arbeitspapier wurden im Rahmen einer Literaturrecherche 30 Quellen identifiziert, die Verfahren zum Umgang mit fehlenden Werten bei qualitativen Daten anhand von Simulationen vergleichen. Der Aufbau des Arbeitspapiers orientiert sich an Rockel (2017, S. 2–13). Im Folgenden wird zunächst der generelle Aufbau von Simulationsstudien beschrieben. Anschließend wird auf die verschiedenen Faktoren, die im Rahmen einer Simulationsstudie variiert werden können, im Einzelnen eingegangen. Das Arbeitspapier schließt mit einem Fazit.

2 Aufbau der untersuchten Simulationsstudien

In den untersuchten Quellen gibt es zwei verschiedene Vorgehensweisen zum Vergleich der MD-Verfahren. Die Unterschiede zwischen diesen beiden werden durch die verwendete Datenbasis verursacht. Bei der einen Vorgehensweise werden reale unvollständige Daten verwendet, während die andere auf vollständigen Daten beruht.

Abbildung 1: Vorgehensweise bei einer realen unvollständigen Datenmatrix (in Anlehnung an Rockel (2017, S. 3))

Die grundsätzliche Vorgehensweise bei der Verwendung realer unvollständiger Daten ist in der Abbildung 1 schematisch dargestellt. Da bei dieser Vorgehensweise die Datenmatrizen bereits unvollständig sind, können die MD-Verfahren direkt auf die Matrizen angewendet und anschließend die Ergebnisse der einzelnen MD-Verfahren miteinander verglichen werden (vgl. z. B. Josse et al. 2012, van der Palm et al. 2016b). Die Vorgehensweise, die auf vollständigen Daten beruht, ist in der Abbildung 2 dargestellt.

Abbildung 2: Vorgehensweise bei einer vollständigen Datenmatrix (in Anlehnung an Rockel (2017, S. 3))

Unter vollständigen Datenmatrizen werden

- simulierte Daten,
- reale vollständig erhobene Daten und
- reale Daten, deren fehlende Werte entweder gelöscht („real komplett“) oder durch Imputation zunächst vervollständigt werden („real imputiert“),

verstanden. Um die MD-Verfahren sinnvoll anwenden zu können, müssen bei diesen Datenmatrizen zunächst Werte gelöscht werden, bevor die MD-Verfahren angewendet werden. Anschließend können die Ergebnisse der MD-Verfahren analysiert werden. Dazu können diese entweder mit bekannten Simulationsparametern oder mit Ergebnissen, basierend auf der vollständigen Datenmatrix, verglichen werden. Die Schritte vom Generieren der Datenmatrix (falls diese simuliert sind) bzw. vom Löschen der Werte (falls reale Daten verwendet werden) bis zur Analyse der Ergebnisse wird in vielen Simulationen mehrmals wiederholt und die Analysen über die Wiederholungen aggregiert. Wie viele Wiederholungen bei der zweiten Vorgehensweise in den untersuchten Quellen verwendet werden, zeigt die Abbildung 3.[1]

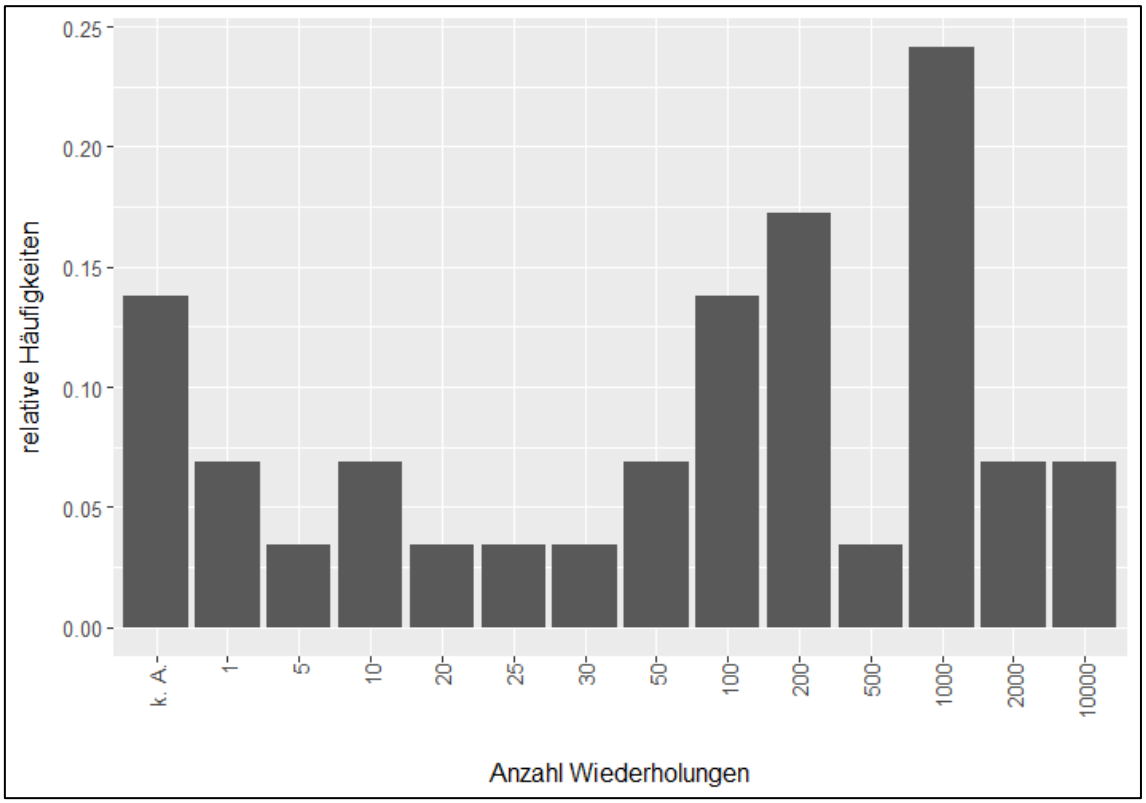

Abbildung 3: Anzahl der Wiederholungen bei vollständigen Daten

[1] Studien, basierend auf real unvollständigen Datenmatrizen, führen keine Wiederholungen zur Erzeugung des zu analysierenden Datenmaterials durch und werden deshalb in der Abbildung nicht berücksichtigt.

Studien, in denen die Anzahl der Wiederholungen nicht explizit angegeben wird und bei denen die Anzahl an Wiederholungen auch nicht durch andere Hinweise ermittelt werden kann, sind in der Abbildung durch „k. A." dargestellt. Anhand der Abbildung ist zu erkennen, dass als Anzahl der Wiederholungen ausschließlich „runde" Zahlen gewählt werden und diese stark variieren. Am häufigsten werden 100, 200 oder 1000 Wiederholungen durchgeführt bzw. die Anzahl der Wiederholungen nicht angegeben. Die Auswahl der Wiederholungsanzahl wird in der Regel nicht begründet. Damit bleibt die Anzahl der Wiederholungen oftmals nicht nur unbegründet, sondern wird stellenweise gar nicht erst angegeben.

2.1 Datenmatrizen

In diesem Abschnitt werden die Beobachtungen zu den Datenmatrizen und Strukturen dieser dargelegt. Hierbei wird untersucht, wie viele Ausprägungen die betrachteten Merkmale besitzen, welche Skalenniveaus vorliegen und welche Arten von Datenmatrizen zur Untersuchung verwendet werden.

Die Anzahl der Ausprägungen (oder auch Anzahl der Kategorien) der verwendeten Merkmale wird für die Auswertung in dichotom, trichotom und mehr als drei Ausprägungen unterschieden. Die genaue Anzahl der betrachteten Ausprägungen ist in der folgenden Tabelle 1 zu sehen.

	Ausprägungen			
	2	3	> 3	k. A.
Quellen	23	13	22	1

Tabelle 1: Anzahl der Ausprägungen

Wie der Tabelle zu entnehmen ist, wird in einer Quelle die Anzahl der verwendeten Kategorien in den Merkmalen nicht angegeben. Dementsprechend ist dieser Fall als „k. A." gekennzeichnet. Falls mehrere Merkmale mit unterschiedlichen Anzahlen an Ausprägungen in einer Quelle betrachtet werden, erfolgt die Erfassung der Ausprägungen der einzelnen Merkmale separat. Wenn z. B. in einer Quelle sowohl dichotome als auch trichotome Merkmale in den Datenmatrizen vorkommen, wird diese Quelle in den Spalten für „2" und „3" Ausprägungen erfasst. Bei Studien, in denen auch quantitative Merkmale in den Datenmatrizen vorkommen, wird für die quantitativen Merkmale stets eine Anzahl an Ausprägungen von mehr als drei unterstellt und dementsprechend erfasst. Am häufigsten werden Merkmale mit genau zwei oder mehr als drei Ausprägungen betrachtet. Der Grund hierfür ist, dass sich einige Quellen zum einen u. a. mit der Güte von MD-Verfahren für

rein binäres Datenmaterial beschäftigen (z. B. Faisal und Tutz 2017) und zum anderen rein trichotomes Datenmaterial hingegen nur in einzelnen Fällen betrachtet wird (z. B. van der Palm et al. 2016a).

Weitere Untersuchungspunkte sind das Skalenniveau der verwendeten Merkmale und die Art der Datenmatrix. In der folgenden Tabelle 2 sind die betrachteten Skalenniveaus sowie die dabei untersuchten Arten von Datenmatrizen zusammengefasst.

		Skalenniveau					
		Binär	Nominal	Ordinal	Gemischt	Unbekannt	Gesamt
Datenmatrix	Real unvollständig	3	3	1	0	1	5
	Real vollständig	6	6	2	1	0	7
	Real komplett	4	2	2	2	1	7
	Real imputiert	1	0	0	0	0	1
	Real unbekannt	1	2	1	1	1	3
	Simuliert	12	6	4	3	1	16
	Resampling	3	1	1	1	1	4
	Gesamt	23	16	8	6	4	30

Tabelle 2: Arten der Datenmatrix und Skalenniveaus der untersuchten Quellen

Die Einteilung der Arten von Datenmatrizen orientiert sich an der Unterteilung, die bereits zu Beginn des Kapitels 2 verwendet wird. Neben den dort beschriebenen Arten von Datenmatrizen sind zusätzlich die Optionen „real unbekannt“ und „Resampling“ in der Tabelle aufgeführt. Bei „real unbekannten“ Daten, handelt es sich um jene Datenmatrizen, für die nur ersichtlich ist, dass sie real erhoben sind, jedoch weitere Angaben zur Vollständigkeit oder eventuellen weiteren Bearbeitungsschritten zur Vervollständigung fehlen. „Resampling“ dagegen beschreibt einen Prozess, bei dem aus einer vollständigen Datenmatrix Stichproben (sogenannte Subsamples) gezogen und diese Subsamples als eigenständige Datenmatrizen in der weiteren Analyse verwendet werden. Die Zeilen- bzw. Spaltensummen in der Tabelle entsprechen meist nicht den Gesamtanzahlen der Arten von Datenmatrizen bzw. der Skalenniveaus, da in vielen Quellen unterschiedliche Datenmatrizen zum Vergleich der dort untersuchten Verfahren herangezogen werden.

In 19 der 30 Quellen wird mindestens eine reale Datenmatrix verwendet, wohingegen in 16 Quellen mindestens eine simulierte Datenmatrix genutzt wird. 9 der 30 Autoren entscheiden sich sowohl für simulierte als auch für reale Datenmatrizen in ihren Untersuchungen, selten verwenden sie die Option des „Resamplings“. Nur in fünf Quellen werden „real unvollständige“ Datenmatrizen verwendet. Eine mögliche Begründung dafür liegt in der erschwerten Beurteilung der MD-Verfahren bei der Verwendung real unvollständiger Da-

tenmatrizen, was u. a. aus der fehlenden Vergleichsmöglichkeit mit zuvor festgelegten Simulationsparametern oder vollständigen Daten resultiert.

Die Skalenniveaus werden, wie für qualitativen Daten typisch, in nominal und ordinal unterteilt (vgl. Bamberg et al. 2017, 6 f.). Da, wie zuvor beschrieben, die dichotomen bzw. binären Daten eine besondere Stellung einnehmen, werden diese in der Tabelle zusätzlich erfasst. Unter „gemischt“ sind alle Quellen zusammengefasst, bei denen in einer Untersuchung mindestens eine Datenmatrix verwendet wird, die sowohl qualitative als auch quantitative Merkmale beinhaltet. Dabei werden die Skalenniveaus der qualitativen Merkmale zusätzlich einzeln in der Tabelle erfasst. Eine weitere Besonderheit ist mit dem Skalenniveau „unbekannt“ gegeben. Hierzu zählen alle Datenmatrizen, bei denen nicht erkennbar ist, welches Skalenniveau in mindestens einem Merkmal vorliegt.

Am häufigsten wird in den Quellen mindestens ein binäres Merkmal verwendet (23 von 30). Diese Quellen haben sich in ihren Untersuchungen jedoch nicht nur auf rein binär skaliertes Datenmaterial beschränkt, sondern auch die Kombination mit nominalen und ordinalen Daten untersucht. Vier von acht Quellen, welche zur Untersuchung u. a. ordinalskalierte Merkmale betrachten, nutzen ausschließlich ordinale Merkmale (z. B. Cugnata und Salini 2017 und Wu et al. 2015). Ferner ist im Inneren der Tabelle 2 zu erkennen, welche Kombinationen aus Art von Datenmatrix und Skalenniveaus in den einzelnen Quellen verwendet werden. Beispielsweise wird in drei Quellen ein real unvollständige Datenmatrix genutzt, in dem mindestens ein Merkmal binär skaliert ist. Am häufigsten wird die Kombination aus simulierten Datenmatrizen mit mindestens einem binären Merkmal verwendet.

2.2 Fehlende Werte

Einen weiteren relevanten Untersuchungspunkt bilden die fehlenden Werte innerhalb der Datenmatrizen. Dazu werden der jeweils vorliegende Ausfallmechanismus, das Ausfallmuster und der Anteil fehlender Werte genauer betrachtet. Die in den Studien auftretenden Kombinationen aus Ausfallmechanismus und Ausfallmuster sind in der folgenden Tabelle 3 zusammengefasst dargestellt. Grundsätzlich werden die Ausfallmechanismen, wie in Kapitel 1 zuvor beschrieben, in MCAR, MAR und MNAR unterteilt.

		Muster		
		Univariat	Multivariat	Gesamt
Mechanismus	MCAR	8	16	23
	MAR	6	12	16
	MNAR	2	2	3
	Real	1	4	5
	Gesamt	11	23	

Tabelle 3: Verwendete Ausfallmechanismen und -muster

Ein weiterer Mechanismus, welcher in der Tabelle aufgeführt wird, ist der „reale“ Ausfallmechanismus. Dieser kann in der Regel nicht genauer bestimmt werden (vgl. Schafer und Graham 2002, S. 152). Dieser Ausfall liegt ausschließlich bei realen Daten vor, die unvollständig erhoben sind. Da die Autoren in ihren Studien mehrere Mechanismen und Muster verwenden können, stimmen die Zeilen- und Spaltensummen fast nie mit den Gesamtzahlen der Ausfallmechanismen und -muster überein. In den Quellen ist die Nutzung des MCAR-Mechanismus am stärksten verbreitet. Er wird in 23 der 30 Quellen genutzt. Der MAR-Mechanismus ist mit 16 von 30 Quellen am zweithäufigsten vertreten. Der MNAR- und der „reale“ Mechanismus hingegen werden nur selten verwendet.

In der Tabelle 3 wird außerdem zwischen univariaten und multivariaten Ausfallmustern unterschieden. Ein univariates Ausfallmuster liegt dann vor, wenn nur in genau einem Merkmal Daten fehlen. Der multivariate Ausfall hingegen dann, wenn mehrere Merkmale vom Ausfall betroffen sind (vgl. van Buuren 2018, S. 105). Aus der Tabelle 3 wird ersichtlich, dass zum einen wesentlich mehr Quellen multivariate als univariate Ausfallmuster betrachten und zum anderen, dass die meisten Quellen einen MCAR- oder MAR-Mechanismus mit einem multivariaten Ausfallmuster verwenden.

Ein weiteres wichtiges Kriterium hinsichtlich des Fehlens der Daten ist, wie eingangs erwähnt, der Anteil dieser. Dazu wird in der folgenden Abbildung 4 dargestellt, wie viele unterschiedliche Ausfallraten in den Quellen angewandt werden.

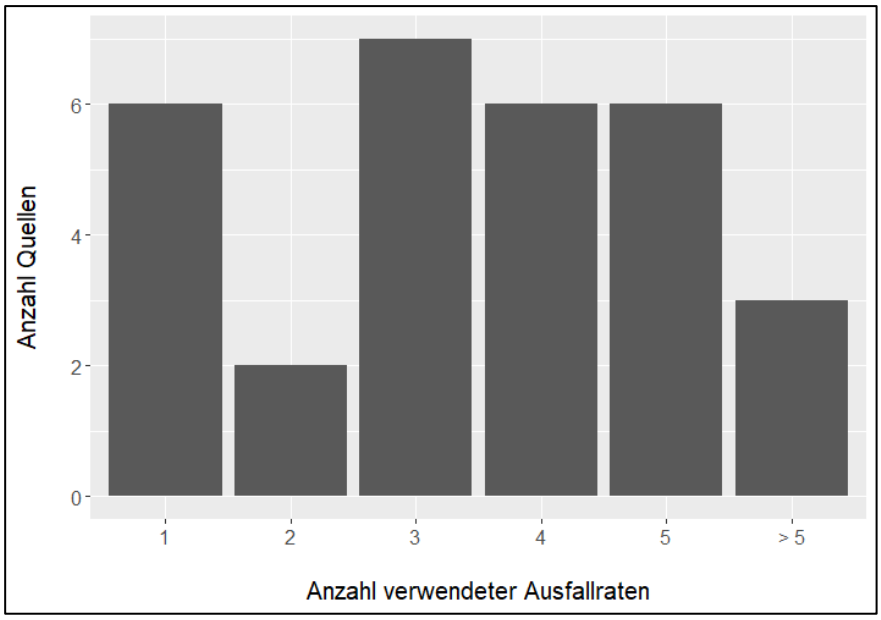

Abbildung 4: Anzahl der verwendeten Ausfallraten

In der Abbildung 4 sind die in den Quellen verwendeten Anzahlen an Ausfallraten dargestellt. Wenn in einer Quelle die Ausfallrate nicht variiert wird, ist die verwendete Anzahl entsprechend eins. Am häufigsten werden in den Quellen eine, drei, vier oder fünf verschiedene Ausfallraten verwendet. Mehr als fünf unterschiedliche Ausfallraten werden nur in drei Quellen verwendet. Neben der Anzahl der verwendeten Ausfallraten in den Quellen wird außerdem festgehalten, in welchen Bereichen die Autoren die gewählten Raten variieren. Dazu sind mittels Boxplots die minimal sowie maximal verwendeten Ausfallraten dargestellt. Diese sind der folgenden Abbildung 5 zu entnehmen.

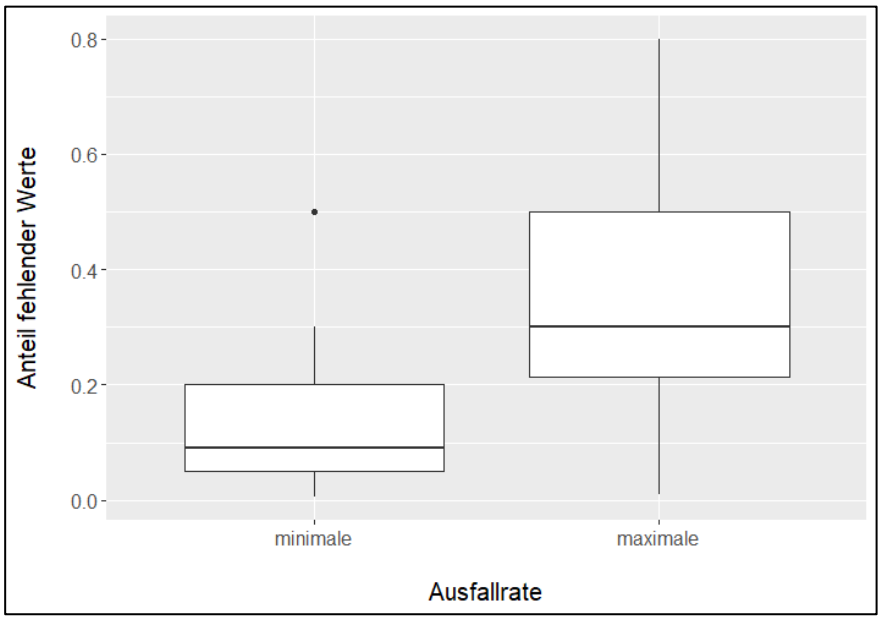

Abbildung 5: Variation des Anteils der fehlenden Werte in den Quellen

Die niedrigste verwendete Ausfallrate beträgt in den Studien 0,5 %. 23 % der Quellen wählen als minimale Ausfallrate weniger als 5 %. Über die Hälfte der Quellen hingegen wählen eine minimale Ausfallrate zwischen 5 % und 20 %. Weitere 23 % wählen eine minimale Ausfallrate von mehr als 20 %, wobei diese maximal 30 % beträgt, mit einer Ausnahme in der als minimaler Anteil fehlender Daten 50 % gewählt wird. Die maximal gewählten Ausfallraten reichen von 0,9 % (Ausfall in einer real unvollständigen Datenmatrix) bis hin zu einem Anteil von 80 % fehlender Daten. 87 % der betrachteten Quellen wählen eine Ausfallrate von maximal 50 %. Nur 13 % der Quellen wählen eine Ausfallrate von über 50 %. Der entscheidende Einfluss auf die Güte der MD-Verfahren, welchen der Anteil fehlender Werte besitzt, erklärt die häufig unterschiedlich und stark variierenden Anzahlen und Anteile fehlender Werte.

2.3 MD-Verfahren

Das Hauptaugenmerk der Simulationsstudien liegt auf den dort untersuchten MD-Verfahren. In diesem Kapitel werden die in den Quellen verwendeten MD-Verfahren betrachtet. Die dazu erfassten Verfahren sind der folgenden Abbildung 6 zu entnehmen. Alle Verfahren, welche mehr als einmal betrachtet werden, sind dort explizit abgebildet. Jene Verfahren, die nur einmal in den Quellen vorkommen, sind unter dem Begriff „sonstige“ zusammengefasst. Bei diesen handelt es sich um sehr spezielle Verfahren, welche nicht weiter von anderen Autoren aufgegriffen werden. Da die Autoren in der Regel mehrere MD-Verfahren betrachten, sind die gesamt beobachteten Häufigkeiten der MD-Verfahren deutlich größer, als die Anzahl der untersuchten Quellen. Der Abbildung 6 ist außerdem zu entnehmen, dass MD-Verfahren, die auf Multipler Imputation basieren oder vollständige Objekte mittels Eliminierungsverfahren erzeugen, mit Abstand am häufigsten verwendet werden.

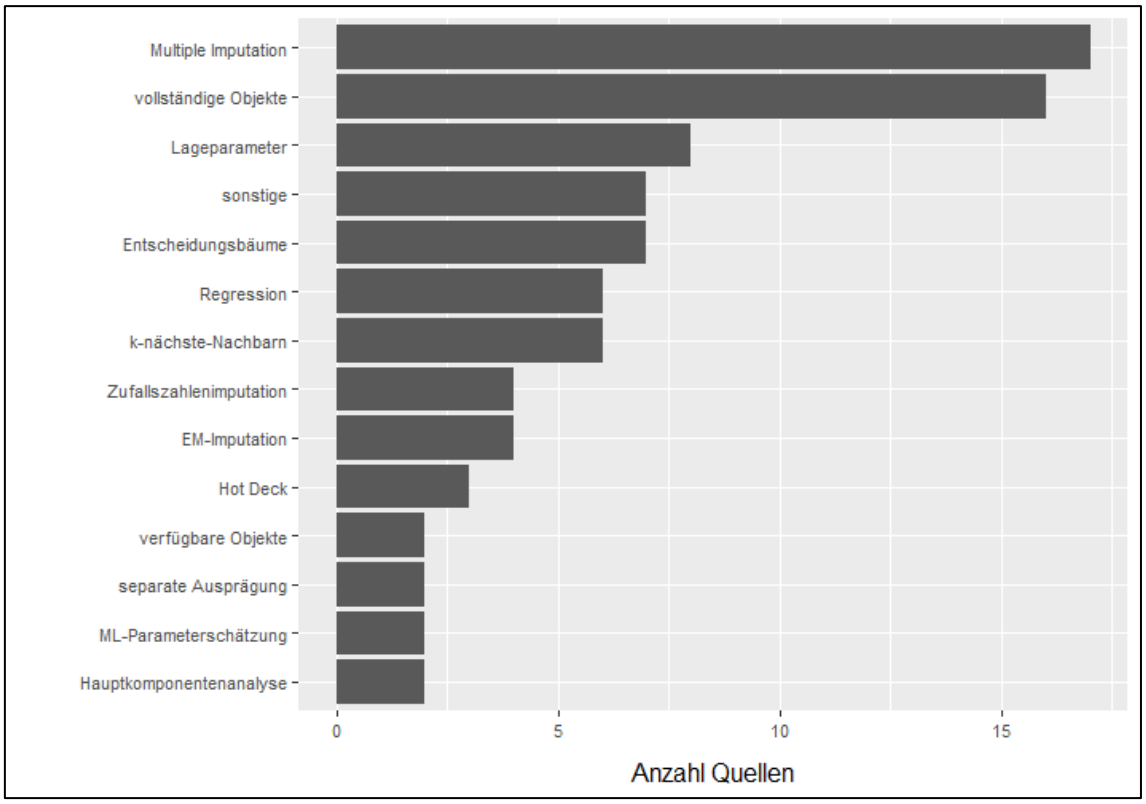

Abbildung 6: Anzahl verwendeter MD-Verfahren

Sowohl die Multiple Imputation als auch die vollständigen Objekte werden in über der Hälfte aller untersuchten Quellen betrachtet. Aufgrund der Einfachheit der Erzeugung vollständiger Objekte mittels Eliminierungsverfahren wird dieses in vielen Studien als Vergleichsbasis herangezogen. Ebenso verhält es sich mit der Imputation von Lageparametern, die aufgrund ihrer Einfachheit ebenfalls oftmals als Vergleichsbasis herangezogen werden.

2.4 Bewertung

In diesem Abschnitt wird die in den Quellen verwendete Art der Bewertung für die untersuchten MD-Verfahren aufgezeigt. Eine entsprechende Übersicht der erfassten Bewertungskriterien ist mit der Abbildung 7 gegeben. Alle Bewertungskriterien, welche nur einmal verwendet werden, sind erneut unter „sonstige“ zusammengefasst. Da die Autoren häufig mehrere Kriterien zur Bewertung der betrachteten Verfahren heranziehen, übersteigen die gesamt beobachteten Häufigkeiten der Bewertungskriterien die Anzahl der betrachteten Quellen. Aus der Abbildung ist ersichtlich, dass überwiegend Regressionsmodelle und die Imputationswerte selbst zur Bewertung herangezogen werden. Im Fall der Regressionsmodelle werden die Verfahren beispielsweise anhand des durch sie hervorgerufenen Bias bei der Schätzung der Regressionskoeffizienten bewertet.

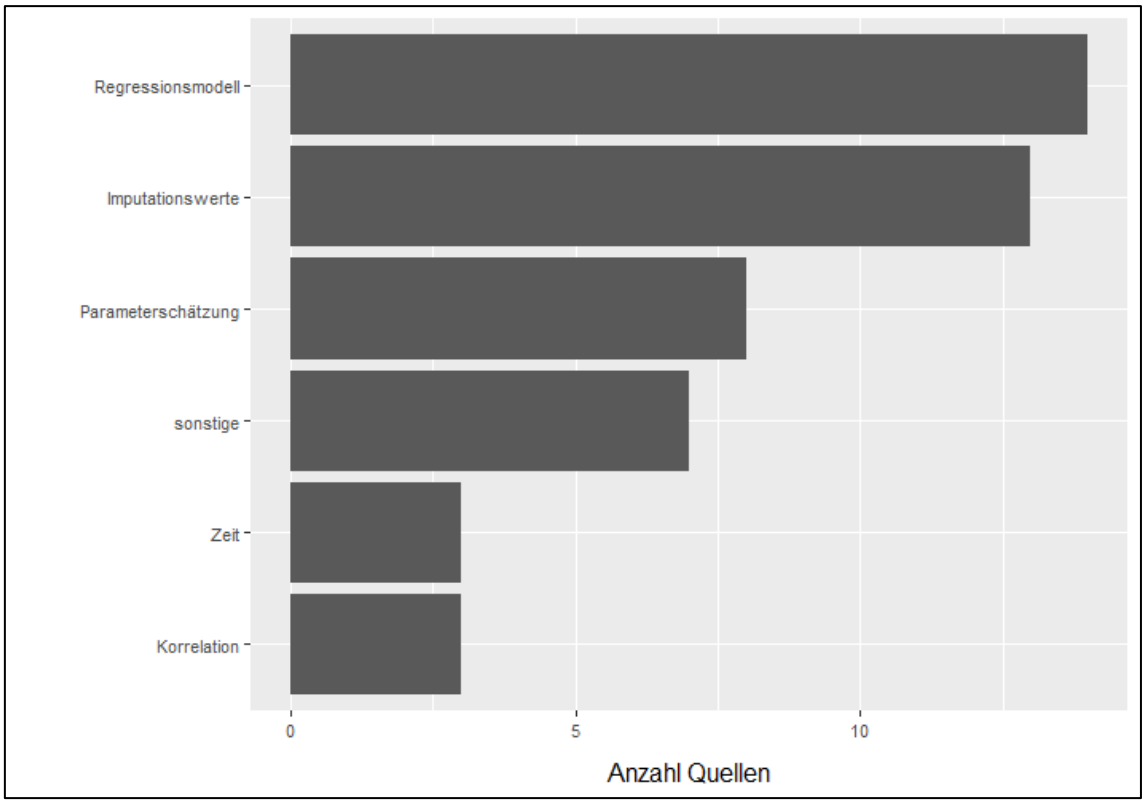

Abbildung 7: Anzahl verwendeter Vergleichskriterien

Im Falle der Imputationswerte wird häufig die Genauigkeit, also der Anteil der korrekt imputierten Werte, im Rahmen der Bewertung betrachtet. Unter Parameterschätzung zählen hier alle Quellen, welche zur Bewertung z. B. Verteilungsparameter bestimmen und anschließend diese mit denen der vervollständigten Datenmatrix vergleichen. Nur wenige Quellen haben zur Beurteilung der MD-Verfahren die erforderliche Zeit zur Imputation und die Korrelation betrachtet.

3 Fazit

Das Ziel des vorliegenden Arbeitspapiers war es, den Aufbau und die Struktur von 30 bereits existierenden Simulationsstudien, speziell für fehlende qualitative Daten, zu betrachten, um daraus Rückschlüsse auf mögliche Forschungslücken ziehen zu können. Dabei zeigte sich, dass die betrachteten Quellen sowohl im Umfang der durchgeführten Simulationen als auch bei deren Beschreibung stark variieren. Besonders auffällig wurde dies, wenn die Autoren beispielsweise zu manchen Simulationsparametern keine Angaben machten. Dieser Umstand beeinträchtigte die Nachvollziehbarkeit einiger Studien.

Außerdem ist zu hinterfragen, ob sich alle betrachteten Studien für einen noch ausstehenden Gütevergleich der dort untersuchten MD-Verfahren eignen. Da, neben den zuvor erwähnten Problemen hinsichtlich der Nachvollziehbarkeit der Simulationen, es auch fraglich ist, inwiefern z. B. Studien mit einer nur geringen Anzahl an durchgeführten Wiederholungen für einen weiteren Vergleich geeignet sind.

Es wurde zudem erkennbar, dass ordinalskalierte Merkmale verhältnismäßig wenig untersucht wurden. Eine Empfehlung basierend auf einer umfangreichen Betrachtung verschiedener MD-Verfahren für unterschiedlich vorliegende Skalenniveaus war außerdem nicht gegeben. Viele Studien verglichen Verfahren für speziell vorliegende Skalenniveaus, doch nur wenige Arbeiten widmeten sich gezielt dem tatsächlichen Unterschied durch die Verwendung verschieden skalierter Merkmale. Aus der Untersuchung ging außerdem hervor, dass in den Studien viele unterschiedliche MD-Verfahren und Bewertungskriterien zum Umgang mit fehlenden qualitativen Daten herangezogen wurden. Dabei konnte kein Verfahren bzw. Vergleichskriterium aufgrund einer besonders häufigen Anwendung hervorstechen.

Mit dem vorliegenden Arbeitspapier konnte festgestellt werden, wie die Autoren ihre Simulationen für fehlende qualitative Daten aufbauen. Aussagen bzgl. der Häufigkeit und Verhältnismäßigkeit in den Beobachtungen sind jedoch immer unter der Prämisse zu sehen, dass zunächst nur 30 Studien betrachtet wurden. Dadurch ist es sinnvoll, unter Berücksichtigung der zuvor benannten Punkte den Umfang der untersuchten Studien zu erweitern und anschließend einen umfangreichen Gütevergleich der MD-Verfahren für fehlende Daten in qualitativen Merkmalen anhand bereits existierender Simulationsstudien durchzuführen. Somit können weitere Forschungslücken identifiziert und die hier gewonnenen Erkenntnisse verifiziert werden.

4 Literaturverzeichnis

Backhaus, Klaus; Blechschmidt, Boris (2009): Fehlende Werte und Datenqualität. In: *Die Betriebswirtschaft* 69 (2), S. 265–287.

Bamberg, Günter; Baur, Franz; Krapp, Michael (2017): Statistik. Eine Einführung für Wirtschafts- und Sozialwissenschaftler. 18., vollständig aktualisierte Auflage.

Bankhofer, Udo (1995): Unvollständige Daten- und Distanzmatrizen in der Multivariaten Datenanalyse. Bergisch Gladbach, Köln: Eul (Quantitative Ökonomie, 64).

Cugnata, Federica; Salini, Silvia (2017): Comparison of alternative imputation methods for ordinal data. In: *Communications in Statistics - Simulation and Computation* 46 (1), S. 315–330. DOI: 10.1080/03610918.2014.963611.

Eekhout, Iris; Boer, Michiel R. de; Twisk, Jos W. R.; Vet, Henrica C. W. de; Heymans, Martijn W. (2012): Missing Data. A Systematic Review of How They Are Reported and Handled. In: *Epidemiology* 23 (5), S. 729–732. DOI: 10.1097/EDE.0b013e3182576cdb.

Faisal, Shahla; Tutz, Gerhard (2017): Nearest Neighbor Imputation for Categorical Databy Weighting of Attributes. Online verfügbar unter https://arxiv.org/pdf/1710.01011.

Ferrari, Pier Alda; Annoni, Paola; Barbiero, Alessandro; Manzi, Giancarlo (2011): An imputation method for categorical variables with application to nonlinear principal component analysis. In: *Computational Statistics & Data Analysis* 55 (7), S. 2410–2420. DOI: 10.1016/j.csda.2011.02.007.

Josse, Julie; Chavent, Marie; Liquet, Benot; Husson, François (2012): Handling Missing Values with Regularized Iterative Multiple Correspondence Analysis. In: *Journal of Classification* 29 (1), S. 91–116. DOI: 10.1007/s00357-012-9097-0.

Lin, Wei-Chao; Tsai, Chih-Fong (2020): Missing value imputation. A review and analysis of the literature (2006–2017). In: *Artificial Intelligence Review* 53, S. 1487–1509. DOI: 10.1007/s10462-019-09709-4.

Little, Roderick J. A.; Rubin, Donald B. (2020): Statistical analysis with missing data. Third edition. Hoboken, NJ: John Wiley and Sons, Inc.; Wiley (Wiley series in probability and statistics).

Rockel, Tobias (2017): Gütevergleich von Imputationsverfahren - eine Analyse existierender Simulationsstudien. Ilmenau (Ilmenauer Beiträge zur Wirtschaftsinformatik). Online verfügbar unter https://nbn-resolving.org/urn:nbn:de:gbv:ilm1-2017200274.

Rubin, Donald B. (1976): Inference and Missing Data. In: *Biometrika* 63 (3), S. 581–592. DOI: 10.1093/biomet/63.3.581.

Schafer, Joseph L.; Graham, John W. (2002): Missing data: Our view of the state of the art. In: *Psychological Methods* 7 (2), S. 147–177. DOI: 10.1037//1082-989X.7.2.147.

van Buuren, Stef (2018): Flexible imputation of missing data. Second edition. Boca Raton, London, New York: Chapman and Hall/CRC (Chapman and Hall/CRC Interdisciplinary statistics series).

van der Palm, Daniël W.; van der Ark, L. Andries; Vermunt, Jeroen K. (2016a): A comparison of incomplete-data methods for categorical data. In: *Statistical methods in medical research* 25 (2), S. 754–774. DOI: 10.1177/0962280212465502.

van der Palm, Daniël W.; van der Ark, L. Andries; Vermunt, Jeroen K. (2016b): Divisive Latent Class Modeling as a Density Estimation Method for Categorical Data. In: *Journal of Classification* 33 (1), S. 52–72. DOI: 10.1007/s00357-016-9195-5.

Wu, Wei; Jia, Fan; Enders, Craig (2015): A Comparison of Imputation Strategies for Ordinal Missing Data on Likert Scale Variables. In: *Multivariate behavioral research* 50 (5), S. 484–503. DOI: 10.1080/00273171.2015.1022644.